

[illegible]

segmenting an input s  
n-segmented language, t  
possible segments  
ce of characters, at le  
ssible segments overlap  
an alternative seq  
ters for at least on  
le segments, the a  
ce of characters fo  
ative segment; and  
multiple syntactic anal  
possible segments  
ative segment, the  
ing in a full syntac  
tilizes and thereby res  
tation of the input se  
ters.

of claim 1 wherein p  
analyses comprises p  
t in a parse conta  
of claim 1 wherein iden  
of characters for a  
ntifying an alternative

- identifying possible segments in the sequence of characters, at least two of the possible segments overlapping each other;

identifying an alternative sequence of characters for at least one of the possible segments, the alternative sequence of characters forming an alternative segment; and

performing multiple syntactic analyses using the possible segments and the alternative segment, the analyses resulting in a full syntactic parse that utilizes and thereby results in a segmentation of the input sequence of characters.

2. The method of claim 1 wherein performing multiple syntactic analyses comprises performing analyses that result in a parse containing the alternative segment.

3. The method of claim 1 wherein identifying an alternative sequence of characters for a possible segment comprises identifying an alternative sequence

of characters that has a different number of characters than the possible segment.

4. The method of claim 3 wherein performing multiple syntactic analyses comprises treating an alternative segment that has a different number of characters than the possible segment for which it is an alternative as if it had the same number of characters as the possible segment for which it is an alternative.

5. The method of claim 1 wherein identifying an alternative sequence of characters comprises performing inflectional morphology on a possible segment.

6. The method of claim 1 wherein identifying an alternative sequence of characters comprises identifying orthographic variations of a possible segment.

7. The method of claim 6 wherein identifying orthographic variations comprises identifying a preferred orthographic form for the possible segment.

8. The method of claim 1 wherein identifying orthographic variations comprises identifying orthographic variants that use a script other than the script of the characters in the possible segment.

Set  
A1

001011-0010250

Figure 1 displays 12 histograms, labeled  $x_1$  through  $x_{12}$ , showing the distribution of the number of non-zero elements in the vector  $x_k$ . The x-axis represents the number of non-zero elements (0 to 10), and the y-axis represents the count (0 to 10). The distributions are roughly bell-shaped and centered around 5, with the peak count increasing from 10 for  $x_1$  to 10 for  $x_{12}$ .

a word breaker that generates a collection of words from the string of characters, the collection of words comprising at least two words that are derived in part from the same character in the string of characters, the word breaker utilizing:

a variants constructor that is used to derive word variants of words found in the string of characters, each word variant being added to the collection of words and having a different sequence of characters than the sequence of characters associated with the word in the string of characters from which it is derived; and

a syntax parser that performs a syntactic analysis using the collection of words produced by the word breaker to produce a syntax parse, the syntax parse

indicating the syntax of the string of characters.

10. The system of claim 9 wherein the variants constructor comprises inflectional morphology rules.

11. The system of claim 10 wherein the inflectional morphology rules are capable of identifying a word's lemma from its inflectional form in the string of characters.

12. The system of claim 9 wherein the variants constructor comprises an orthographic variants structure that indicates the orthographic variants of words found in the string of characters.

13. The system of claim 9 wherein at least one word variant has a different number of characters than the word from which it is derived.

14. The system of claim 9 wherein at least one word variant includes a character that is not present in the string of characters.

15. A computer-readable medium having computer-executable instructions for performing steps comprising:

See  
A1

00000000-10100

[illegible]

identifying a variant for a first group of characters in the sequence of characters, the variant containing a different collection of characters than the collection of characters in the first group of characters;

identifying a second group of characters in  
the sequence of characters that  
overlaps the first group of characters;  
and

performing a syntactic analysis using the variant and the second group of characters to produce a syntactic parse, the syntactic parse containing either the variant or the second group of characters.

16. The computer-readable medium of claim 15 wherein identifying a variant for a first group of characters comprises identifying a variant that has a character that is not present in the first group of characters.

17. The computer-readable medium of claim 15 wherein identifying a variant for a first group of characters comprises identifying a variant that has a



~~computer-readable  
giving an ortho  
satisfying a pr  
  
computer-readable  
giving an ortho  
satisfying a var  
of a different  
in the first~~

[illegible]